B-109E
c.2.

Statistics Canada Statistique Canada

Ottawa

A CLOSER LOOK AT THE NEAREST NEIGHBOUR/HOT DECK IMPUTATION METHODS: AN EMPIRICAL STUDY

John G. Kovar
I.A.S.M.D.
February 1982

1. SUMMARY

The present study is an empirical comparison of the nearest neighbour and the hot deck imputation methods for partial nonresponse using agriculture survey data. The reader is referred to the author's previous report (Kovar, 1981) for a description and a theoretical comparison of the two imputation strategies. The nearest neighbour and the hot deck imputation methods were compared using the 1981 FES (Farm Enumerative Survey) data. Random nonresponse was introduced in three expense sections of the questionnaire in twenty percent of all Saskatchewan records. The "dirty" file thus created, was in turn "cleaned" up by both imputation procedures and passed through the existing tabulation and "top forty contributors to the estimate" programs. The results reveal little or no difference between the estimates, with the hot deck generally overestimating the "true" value slightly. (The "true value" refers here to the estimate from the original data file.) On the other hand, even though the estimated coefficients of variation are very close to their corresponding "true" values (as estimated from the original data file), the hot deck estimates of the CV's are higher or equal to the nearest neighbour estimates in all cases. (On the average, the hot deck procedure tends to inflate the coefficient of variation while the nearest neighbour procedure tends to underestimate it.) Finally, it is noted that the hot deck method seems to create more outliers than does the nearest neighbour.

2. METHODS

The 1981 FES post imputed file was chosen as the "true" underlying population for the purposes of this exercise. It was realized at that time that the file contains duplicate fields, as it was imputed by a hot deck procedure*, and thus will not be an ideal test file. The situation could have been rectified by removing all imputed records (or at least those where the imputations affected the fields of interest) but this would have necessitated revision of the raising factor file needed by the tabulation program. The laboriousness of modifying the raising factors was deemed to far outweigh the possible danger of introducing a systematic bias to the study due to duplicates, especially when the comparison in question is of a scale as small as that presented here.

Secondly, three expense sections of the questionnaire were selected as the test sections for imputation. The final selection was a function of several factors. Since this test is in part a feasibility study of a Farm Energy Use Survey (FEUS) imputation system, an attempt was made to select fields or sections resembling those of the FEUS. Hence fuel expense questions and questions pertaining to expenditures rather than crops and livestock were used. Equally important was the

*In the three sections considered, such duplication is of the order of less than 10%.

consideration (necessity) to select imputation paths which stand, more or less, on their own in the hot deck imputation program. (That is, those that do not trigger a chain of other imputations.) Finally an effort was made to select imputation paths which vary as widely as the hot deck approach allows in their donor selection criteria. (That is, matching on one or more continuous or categorical variable.) The final choice included section H12 (Electricity, Telephone and Heating Fuel Expenses), H15 (Custom Work and Machine Hire Expenses) and H11 (Interest on Farm Loans, Credit and Mortgages).

Due to size considerations, only records in the province of Saskatchewan were considered for the test purposes. Of the total 3644 records (excluding community pastures), three independent simple random samples of approximately 729 records each (20%) were selected. The records in samples one, two or three were flagged as nonrespondents to sections H12, H15 and H11 respectively, creating a total of 1776 records with one or more nonresponse. Operationally, this was accomplished by changing the section refusal boxes (fields 830(H12), 760 (H15) and 820(H11)) from zero to nine. This will in turn trigger imputations 169, 135 and 163 in the hot deck program respectively and will affect a total of seven data fields in the three sections.

Thirdly, the three hot deck imputations were mimicked using the Numerical Edit and Imputation system (Sande, 1979) imbedded it the P-STAT statistical software package (Buhler and Buhler, 1979). Namely, imputation 169 was imitated by matching on one continuous variable (Total Agriculture Receipts, field 901). Imputation 135 used two continuous matching variables (Total Land Operated in the Previous Year, field 222, in addition to that above) and imputation 163 used a categorical variable (soil zone, which takes on four values in Saskatchewan) in addition to a continous variable (field 901). (See Appendix B) Note that an attempt was made only to duplicate the existing imputation procedure by the nearest neighbour system, rather than better it. (The strategy of using the nearest neighbour procedure to its full potential, for example by using more matching variables, was discarded as it would unnecessarily introduce an extra factor to the study.) The final results therefore only indicate how the nearest neighbour procedure performs in a hot deck setting.

3. RESULTS

Two imputed files were created (one by the hot deck and one by the nearest neighbour procedure) and subsequently passed through existing programs to evaluate the impact of the imputation method on the estimates and their CV's at the Saskatchewan level. As well, the top forty contributors to the estimate were examined in order to compare the procedures in terms of their susceptibility to create outliers.

Table I. shows the estimates and their CV's before imputation ("true") as well as after imputation via nearest neighbour procedure (NN) and hot deck procedure (HD). All results are shown for the seven affected fields (760's in section H15, 820's in H11 and 830's

in H12) at the Saskatchewan level. Note that in all cases the estimated CV's* are higher (or equal) in the case of the hot deck imputation than those of the nearest neighbour imputation and in all but one case than the true CV's. On the other hand, in all but two cases the nearest neighbour CV's were lower than the true values. This trend is reflected in the average CV over the seven fields (viz. $3.44 < 3.63 < 3.74$), though the effect is less pronounced when field 762 is disregarded (viz. $3.05 < 3.11 < 3.17$). (A possible explanation of this phenomenon lies in the fact that hot deck imputation methods introduce a new component to the variance term due to the random selection (different with every ordering of the file) of the donor records. On the other hand a nearest neighbour procedure is virtually deterministic in the sense that the same donor will be selected for each given recipient regardless of the order of presentation.)

Table II. shows the item estimates after imputation in terms of the percentage of the corresponding "true" values. Note that while the hot deck overestimates the "true" value on the average by about 1%, the absolute percent deviations are comparable to those of the nearest neighbour (2.09% *vs.* 2.14%).

Finally, Table III. presents the number of new records introduced into the set of the top forty (Table IIIA) and twenty (Table IIIB) contributors to the estimate from those records which were not part of the original ("true") top forty contributors for each of the seven items in question. As well, the rank of the highest new contributor is presented (Table IIIC). As expected, the hot deck method has a tendency to create more outliers than the nearest neighbour.

4. COMMENTS

Overall, the nearest neighbour procedure fared quite well as compared to the hot deck. It must be remembered that the nearest neighbour was supposed to only duplicate the hot deck results and it did so effectively. It is felt that the option of using more matching variables than were used in this study makes the nearest neighbour a very attractive alternative. In the same vein, there seems to be no reason why the nearest neighbour imputation procedure should not be used for the upcoming FEUS. However, some other points should be investigated before a final decision is reached vis-à-vis the IAS (the Integrated Agriculture Survey).

First, there is the question of operational cost (computer time and core usage). While the nearest neighbour performed satisfactorily in terms of time, there are limitations of size in the present system when it comes to creating the actual search tree. While this is not an insurmountable obstacle, it must be evaluated further. A strict comparison of both time and core of the nearest neighbour to the hot deck is not possible at this time, due to the fact that existing pro-

*CV's are estimated using replicates built into the sample design.

grams (written for other purposes) were used during the study. (Programs such as the hot deck imputation module have a high set up and overhead time associated with them and it would be unfair to compare their running time and space requirements to that of a small module such as the nearest neighbour package used here.) These technical aspects of both systems deserve further attention.

Secondly, it is unclear whether some of the results (and/or conclusions) are in part affected by the duplicates in the original file. (It could perhaps be argued that nearest neighbour is better equipped to locate such duplicates as donors and hence reduce the CV's—as was noted earlier.) Perhaps another true population could be defined using the 1981 census data file. It must be noted, however, that such an exercise would necessitate the creation of additional software (hot deck and tabulation programs). Along the same lines, the hypothesis of additional variance component being introduced by hot deck could be tested by rearranging the file in a different order and rerunning the hot deck imputation program. If the hypothesis is correct, averaging out the imputed values over the two runs should bring the CV's closer to their "true" values. Finally, it could prove useful to investigate the frequency distributions of the "true" and post imputed populations, as well as the effects of nonrandom nonresponse.

In conclusion, the nearest neighbour imputation procedure appears to offer a very attractive alternative to the hot deck package at this time. Additional comparison of the two methods is, however, recommended.

5. REFERENCES

Buhler, S. and Buhler, R. (1979). P-STAT 78 User's Manual. Version 4. Princeton University

Kovar, J. (1981). Edit and Imputation Package for the Integrated Agriculture Survey. IASMD Working Paper.

Sande, G. (1979). Numerical Edit and Imputation. Invited paper at the 42nd session of the International Statistical Institute, Manila, Phillipines, Dec. 4-14.

TABLE I. Estimates and their CV's before and after imputation for the province of Saskatchewan

Field	Estimate ($) and CV (%)		
	Post Imputed NN	Post Imputed HD	"True" Original Estimate
762: Total Custom Work Expenses ...	59,539,594 (5.8)	60,936,650 (7.2)	63,562,059 (6.4)
822: Total Interest Expenses	275,208,569 (3.2)	271,534,550 (3.2)	265,771,451 (3.0)
824: Mortgage Interest Expenses	275,611,542 (3.2)	256,495,198 (3.3)	250,500,948 (3.1)
832: Telephone Expenses	13,553,077 (2.8)	13,874,437 (3.0)	13,733,438 (3.0)
833: Electricity Expenses	31,253,043 (2.5)	31,427,754 (2.7)	31,193,502 (2.7)
834: Fuel Expenses	21,904,409 (4.1)	22,336,209 (4.2)	21,746,849 (4.2)
835: Total of above 3 Expenses	66,714,662 (2.5)	67,638,401 (2.6)	66,676,850 (2.7)
Average CV	(3.44)	(3.74)	(3.63)

TABLE II. Post imputed estimates expressed as percent of the "true" value

Field	% of "true" value (% absolute deviation)			
	NN		HD	
762: Total Custom Work Expenses	93.7	(6.3)	95.9	(4.1)
822: Total Interest Expenses	103.6	(3.6)	102.2	(2.2)
824: Mortgage Interest Expenses	102.8	(2.8)	102.4	(2.4)
832: Telephone Expenses	98.7	(1.3)	101.0	(1.0)
833: Electricity Expenses	100.2	(0.2)	100.8	(0.8)
834: Fuel Expenses	100.7	(0.7)	102.7	(2.7)
835: Total of above 3 Expenses	100.1	(0.1)	101.4	(1.4)
Average	99.97	(2.14)	100.91	(2.09)

TABLE III. Outlier creation. A comparison

A. Number of farms introduced to the top forty contributors to the estimate from beyond the "true" top forty.

FIELD	NN	HD
762	10	9
822	8	12
824	5	12
832	7	7
833	9	10
834	10	13
835	8	8
Average	8.14	10.14

B. Number of farms introduced to the top twenty contributors to the estimate from beyond the "true" top forty.

FIELD	NN	HD
762	1	2
822	3	5
824	4	6
832	3	4
833	6	5
834	5	8
835	4	6
Average	3.71	5.14

C. Position (rank in the top forty) of the highest new contributor introduced by imputation.

FIELD	NN	HD
762	7	5
822*	1	1
824*	1	1
832	6	3
833	3	3
834	6	3
835	10	5
Average	4.86	3.00

*This contributor is the same for both imputation procedures and results due to a high raising factor rather than an unduly high donor value.

APPENDIX A

Some of the existing files created during the above described study are listed below for future access.

SASKFES.DIRTY (vol = 024293) - 20% nonresponse in Saskatchewan records in three fields . Contains all FES fields and all FES records.
(2400 long, 9761 records)

SASKFES.ALL (vol = 046179) - 20% nonresponse introduced to three fields. Saskatchewan records only. Contains only those fields needed for imputations in the study.
(119 long, 3644 records - subset of SASKFES.DIRTY)

SASKFES.CLEANED.BYNN (vol = 062011) - Same layout as SASKFES.DIRTY. Post imputed file using nearest neighbour.

SASKFES.ALL.CLEAN (vol = 050922) - Same layout as SASKFES.ALL. Post imputed using nearest neighbour.
(Subset of SASKFES.CLEANED.BYNN)

SASKFES.CLEANED.BYHD (vol = 011612) - Same as SASKFES.CLEANED.BYNN but using hot deck procedures.

APPENDIX B

Overview of recipient-donor matching criteria
under two imputation methods

Imputation	Condition	Affected Fields	HD Donor Selection Criteria *	NN Donor Selection Criteria
135 Custom Work Expenses Section Refused	Field 760 = 9 (complete section refusal)	762	Select 1 of 9 matrices based on 3 ranges of income (901) and 3 ranges of total land (222)	Match as closely as possible on income (901) and total land (222)
163 Interest Expenses Section Refused	Field 820 = 9 (complete section refusal)	822 824	Select 1 of 6 matrices based on 6 ranges of income (901)	Match as closely as possible on income (901), and exactly on soil zone
169 Telephone, Electricity, and Fuel Expenses Section Refused	Field 830 = 9 (complete section refusal)	832 833 834 8[illegible]	Select 1 of 6 matrices based on 6 ranges of income (901)	Match as closely as possible on income (901)

*The hot deck procedure also splits all matrices by soil zone i.e. it matches implicitely on soil zone for every imputation

B-109F
c.2.

Statistics Canada / Statistique Canada
Ottawa, Canada

IMPUTATION PAR LA MÉTHODE DU PLUS PROCHE VOISIN ET LA MÉTHODE "HOT DECK": ÉTUDE EMPIRIQUE

John G. Kovar
Division des méthodes
d'enquête-institutions
et agriculture
février 1982

Canada

1. RESUME

Cette étude présente une comparaison empirique de l'imputation par la méthode du plus proche voisin et la méthode dite du "hot deck" dans les cas de non-réponse partielle à une enquête sur l'agriculture. Le lecteur peut trouver une description et une comparaison théorique de ces deux techniques d'imputation dans un document précédent du même auteur (Kovar, 1981). On compare ici l'imputation par la méthode du plus proche voisin et la méthode du "hot deck" à partir des données de l'EEA (enquête sur les exploitations agricoles) de 1981. Les données de trois sections du questionnaire, qui portaient sur les dépenses, ont été converties au hasard en éléments de non-réponse dans vingt pour cent de tous les enregistrements de la Saskatchewan. Le fichier "brut" ainsi créé a ensuite été "épuré" à l'aide des deux techniques d'imputation susmentionnées et traité par les programmes existants de totalisation et de repérage des "quarante plus grandes valeurs qui entrent dans le calcul d'un estimé". Les résultats ne révèlent presqu'aucune différence entre les estimés, quoique, en général, la méthode du "hot deck" surévalue très légèrement les "vraies" valeurs. (Les "vraies" valeurs correspondent ici aux estimés calculés à partir du premier fichier de données.) En revanche, bien que les coefficients de variation estimés soient très proches des "vraies" valeurs (estimées avec les données du fichier original), les CV estimés par la méthode du "hot deck" sont supérieurs ou égaux aux estimés produits à l'aide de la méthode du plus proche voisin. (Dans l'ensemble, la technique du "hot deck" semble créer plus de valeurs extrêmes que la méthode du plus proche voisin.

2. METHODES

Tout d'abord, le fichier imputé de l'EEA de 1981 a été choisi comme la "vraie" population pour les besoins de cet exercice. Au moment de prendre cette décision, on s'est rendu compte que les données de certains champs de ce fichier étaient des copies d'autres unités, étant donné qu'un traitement d'imputation par la méthode du "hot deck" avait été effectué*. Pour cette raison, ce fichier ne semblait pas idéal pour l'essai proposé. Il aurait été possible de remédier à cette situation en supprimant tous les enregistrements imputés (ou, du moins, les enregistrements sur lesquels l'imputation avait un effet sur les champs visés par l'expérience), mais il aurait alors fallu corriger le fichier des facteurs d'accroissement qu'utilise le programme de totalisation. Le travail nécessaire pour cette révision des facteurs d'accroissement a été jugé beaucoup trop grand en comparaison des conséquences possibles d'un biais systématique engendré par les entrées copiées, d'autant plus que la comparaison allait se faire sur une échelle aussi petite que dans le cas présent.

---

* Dans les trois sections retenues pour cette étude, les données ainsi imputées représentent moins de 10% de l'ensemble des renseignements.

Deuxièmement, trois sections du questionnaire qui portaient sur les dépenses ont été sélectionnées pour l'examen des effets de l'imputation. Ce choix final a été fondé sur plusieurs critères. Comme cette expérience est en partie une étude de faisabilité d'un système d'imputation pour l'enquête sur la consommation d'énergie des exploitations agricoles (ECEEA), on s'est efforcé de prendre des champs ou des sections qui ressemblent à celles de l'ECEEA. Ainsi, on a retenu des questions concernant les dépenses consacrées au carburant et d'autres types de dépenses plutôt que des renseignements sur les cultures et le bétail. Il était également important (nécessaire) de choisir des données dont l'algorithme d'imputation est plus ou moins direct dans le programme du "hot deck" (c'est-à-dire un algorithme qui n'exige pas l'imputation de toute une série d'autres données). Enfin, on s'est appliqué à choisir des données pour lesquelles les critères de sélection d'une valeur imputée sont aussi variés que possible dans la méthode du "hot deck" (c'est-à-dire des données appariées avec une ou plusieurs variables continues ou qualitatives). Le choix final comprenait les sections H12 (dépenses engagées pour l'électricité, le téléphone et le chauffage), H15 (dépenses engagées pour le travail sur commande et la location de machines) et H11 (intérêts sur les prêts et les crédits agricoles et les hypothèques).

A cause de certains facteurs concernant la taille de l'échantillon, on a retenu seulement les enregistrements de la province de Saskatchewan pour les besoins de cette expérience. A partir de cet ensemble de 3,644 enregistrements (qui exclut les pâturages communautaires), on a constitué trois échantillons aléatoires simples indépendants qui comptent chacun environ 729 enregistrements (soit 20% du total). Dans le premier, le deuxième et le troisième échantillon, les cas de non-réponse aux sections H12, H15 et H11 respectivement ont été codés, ce qui à donné un total de 1,776 enregistrements contenant un élément de non-réponse ou plus. Cette opération de codage consiste en réalité à remplacer le code 0 par le code 9 à la case indiquant un refus de répondre à la section concernée (champs 830 (H12), 760 (H15) et 820 (H11)). Ce changement a pour effet de déclencher respectivement les imputations 169, 135 et 163 dans le programme du "hot deck" et, en tout, sept champs de données des trois sections choisies sont modifiées.

Troisièmement, trois traitements d'imputation par la méthode du "hot deck" ont été simulés au moyen du système de contrôle et d'imputation numérique (Sande, 1979) qui fait partie du logiciel statistique P-STAT (Buhler et Buhler, 1979). L'imputation 169 a été simulée par un appariement avec une variable continue (recettes agricoles totales, champ 901). L'imputation 135 s'est faite par un appariement avec deux variables continues (la superficie totale des terres exploitées l'année précédente, champ 222, et la variable susmentionnée), tandis que l'imputation 163 a été exécutée à l'aide d'une variable qualitative (la zone pédologique, dont il y a quatre catégories en Saskatchewan) et d'une variable continue (champ 901). (Voir l'annexe B.) Notons qu'on a seulement voulu reproduire la méthode actuelle d'imputation en appliquant

la technique du plus proche voisin et non l'améliorer. (On a décidé de ne pas essayer de tirer le plein rendement de la technique du plus proche voisin, ce qu'on aurait pu accomplir, par exemple, en utilisant plus de variables d'appariement, parce que cela aurait introduit un facteur additionnel mais non nécessaire dans la présente étude.) Les résultats finals indiquent donc seulement comment la technique du plus proche voisin fonctionne quand elle reproduit la méthode du "hot deck".

3. RESULTATS

Deux fichiers imputés ont été créés (un à l'aide de la méthode du "hot deck" et un par la technique du plus proche voisin) et ont ensuite été traités par les programmes courants pour évaluer l'effet de la méthode d'imputation sur les estimations et les CV calculés pour la Saskatchewan. On a également examiné les quarante plus grandes valeurs qui entrent dans le calcul des estimations afin de comparer le nombre de valeurs extrêmes que créent ces deux méthodes d'imputation.

Le tableau I présente les estimations et les CV calculés avant l'imputation (les "vraies" valeurs) ainsi que les chiffres obtenus après l'imputation par la technique du plus proche voisin et la méthode du "hot deck". Tous les résultats sont indiqués au niveau provincial pour les sept champs modifiées (champ 762 de la section H15, champs 822 et 824 de la section H11 et les champs 832 à 835 de la section H12). Notons que les CV* estimés après imputation par la méthode du "hot deck" sont supérieurs (ou égaux) à chacun des CV calculés après imputation par la technique du plus proche voisin et à tous les vrais CV sauf un. Par contre, tous les CV obtenus après imputation par la technique du plus proche voisin, sauf deux, sont plus faibles que les vraies valeurs. Ces résultats se répercutent sur le CV moyen des sept champs modifiés ($3.44 < 3.63 < 3.74$), quoique la fourchette des CV moyens soit moins étendue si on exclut le champ 762 ($3.05 < 3.11 < 3.17$). Une explication possible de ce phénomène a trait au fait que l'imputation par la méthode du "hot deck" introduit une nouvelle composante dans la variance à cause de la sélection aléatoire (qui produit un résultat différent quand le fichier n'est pas dans le même ordre) des enregistrements qui contiennent les valeurs d'imputation. En revanche, la technique du plus proche voisin est presque déterministe en ce sens que le même enregistrement sera choisi comme source de valeurs d'imputation pour un enregistrement donné indépendamment de l'ordre à l'intérieur du fichier.

Le tableau II exprime chacune des valeurs estimées après imputation en pourcentage de la "vraie" valeur correspondante. On constate que, même si l'imputation par la méthode du "hot deck" conduit à une surestimation moyenne d'environ 1% par rapport aux "vraies" valeurs, les écarts absolus en pourcentage sont comparables à ceux obtenus après imputation par la technique du plus proche voisin (2.09% contre 2.14%).

---

* Les CV sont estimés à partir d'échantillons répétés à l'intérieur du plan d'échantillonnage.

Enfin, le tableau III présente le nombre de nouveaux enregistrements ajoutés à la liste des quarante (tableau IIIA) et des vingt (tableau IIIB) plus grandes valeurs qui entrent dans le calcul des estimés. Ces enregistrements ne faisaient pas partie des quarante plus grandes valeurs qui entrent dans l'estimation de chacune des sept "vraies" valeurs. On trouve aussi au tableau IIIC le rang occupé par la plus grande nouvelle valeur introduite après l'imputation dans le calcul des estimations. Comme prévu, la méthode du "hot deck" crée généralement plus de valeurs extrêmes que celle du plus proche voisin.

4. OBSERVATIONS

Dans l'ensemble, les résultats de la technique du plus proche voisin sont assez bons en comparaison de ceux de la méthode du "hot deck". On doit se rappeler que l'imputation par la technique du plus proche voisin ne visait qu'à reproduire les résultats de l'imputation par la méthode du "hot deck" et que la technique du plus proche voisin s'applique très bien à cette opération. La possibilité d'utiliser plus de variables d'appariement que dans cette étude fait de la technique du plus proche voisin une méthode très prometteuse. En outre, il ne semble pas y avoir de raison de ne pas utiliser l'imputation par la méthode du plus proche voisin dans la prochaine EUEEA. Il faudrait toutefois examiner d'autres aspects avant de prendre une décision finale au sujet de l'enquête nationale sur les fermes.

Premièrement, il faut penser au coût du traitement (temps machine et espace mémoire). Bien que le temps machine requis par la méthode du plus proche voisin soit acceptable, le système actuel impose des contraintes d'espace à la création de l'arbre de recherche. Cet obstacle n'est pas insurmontable, mais il faut l'examiner davantage. Il est impossible à l'heure actuelle de comparer rigoureusement le temps machine et l'espace mémoire requis par les méthodes du plus proche voisin et du "hot deck" parce qu'on a utilisé des programmes qui existaient déjà (et qui avaient été mis au point pour d'autres besoins) pour la présente étude. (Des programmes tels que le module d'imputation par la méthode du "hot deck" requièrent un temps de préparation considérable et nécessitent beaucoup de temps système. Il ne serait pas juste de comparer le temps d'exécution et l'espace nécessaires pour ce genre de programme avec les ressources qu'exige un petit module comme celui dont on s'est servi ici pour l'imputation par la technique du plus proche voisin.) Il faudrait pousser davantage l'étude de ces aspects techniques des deux systèmes d'imputation.

Deuxièmement, il n'est pas évident que certains résultats (et/ou conclusions) sont en partie attribuables au fait que le fichier d'origine contenaient des champs de données copiées. (On pourrait soutenir que la méthode du plus proche voisin est plus axée vers le choix de ces copies comme valeurs d'imputation que le "hot deck", ce qui réduirait les CV -- tel qu'il a été mentionné plus haut.) Il serait peut-être possible de définir une autre vraie population à partir du fichier des données du recensement de 1981. Il faut cependant souligner que cette démarche

exigerait la création de logiciels supplémentaires (programmes de "hot deck" et de totalisation). Il serait aussi possible de vérifier l'hypothèse selon laquelle la méthode du "hot deck" introduit une composante additionnelle de variance. A cette fin, on pourrait changer l'ordre à l'intérieur du fichier et exécuter de nouveau le programme d'imputation par la méthode du "hot deck". Si cette hypothèse est juste, les CV devraient se rapprocher des "vrais" résultats lorsqu'on calcule les estimations à partir des moyennes de chacune des valeurs imputées par la méthode du "hot deck" à deux reprises. Enfin il pourrait être utile d'étudier la distribution de fréquence de la "vraie" population et celle des populations imputées a posteriori, de même que les effets de l'introduction non aléatoire d'éléments de non-réponse.

En conclusion, l'imputation par la technique du plus proche voisin semble actuellement être une solution très avantageuse en comparaison de la méthode du "hot deck". On recommande toutefois que d'autres comparaisons de ces deux méthodes soient effectuées.

5. BIBLIOGRAPHIE

Buhler, S. et Buhler, R. (1979). P-STAT 78 User's Manual: Version 4. Princeton University.

Kovar, J. (1981). Edit and Imputation Package for the Integrated Agriculture Survey. Document de travail de la Division des méthodes d'enquête-institutions et agriculture.

Sande, G. (1979). Numerical Edit and Imputation. Document rédigé en prévision de la 42e séance de l'Institut international de statistique, du 4 au 14 décembre à Manille aux Philippines.

TABLEAU I. Valeurs des estimés et des CV calculés à partir des données de la province de Saskatchewan avant et après l'imputation

CHAMP	Estimations (en dollars) et CV (en pourcentage)		
	Fichier de données imputées par la méthode du plus proche voisin	Fichier de données imputées par la méthode du "hot deck"	"Vraies" valeurs calculées à l'origine
762 : Dépenses totales, travail sur commande	59,539,594 (5.8)	60,936,650 (7.2)	63,562,059 (6.4)
822 : Dépenses totales, intérêts	275,208,569 (3.2)	271,534,550 (3.2)	265,771,451 (3.0)
824 : Dépenses, intérêts Hypothécaires	275,611,542 (3.2)	256,495,198 (3.3)	250,500,948 (3.1)
832 : Dépenses, téléphone	13,553,077 (2.8)	13,874,437 (3.0)	13,733,438 (3.0)
833 : Dépenses, électricité	31,253,043 (2.5)	31,427,754 (2.7)	31,193,502 (2.7)
834 : Dépenses, carburant	21,904,409 (4.1)	22,336,209 (4.2)	21,746,849 (4.2)
835 : Total de ces trois dépenses	66,714,662 (2.5)	67,638,401 (2.6)	66,676,850 (2.7)
CV moyen	(3.44)	(3.74)	(3.63)

TABLEAU II. Estimés calculés après imputation exprimés en pourcentages des "vraies" valeurs

CHAMP	Pourcentage de la "vraie" valeur (écart absolu en pourcentage)			
	Méthode du plus proche voisin		Méthode du "hot deck"	
762 : Dépenses totales, travail sur commande	93.7	(6.3)	95.9	(4.1)
822 : Dépenses totales, intérêts	103.6	(3.6)	102.2	(2.2)
824 : Dépenses, intérêts hypothécaires	102.8	(2.8)	102.4	(2.4)
832 : Dépenses, téléphone	98.7	(1.3)	101.0	(1.0)
833 : Dépenses, électricité ..	100.2	(0.2)	100.8	(0.8)
834 : Dépenses, carburant	100.7	(0.7)	102.7	(2.7)
835 : Total de ces trois dépenses	100.1	(0.1)	101.4	(1.4)
Moyenne	99.97	(2.14)	100.91	(2.09)

TABLEAU III. Création de valeurs extrêmes : comparaison des deux méthodes d'imputation

A. Nombre d'exploitations agricoles auxquelles correspondent les quarante plus grandes valeurs qui entrent dans le calcul des estimations mais qui ne faisaient pas partie de la "vraie" liste des quarante plus grandes valeurs.

CHAMP	METHODE DU PLUS PROCHE VOISIN	METHODE DU "HOT DECK"
762	10	9
822	8	12
824	5	12
832	7	7
833	9	10
834	10	13
835	8	8
Moyenne	8.14	10.14

B. Nombre d'exploitationsagricoles auxquelles correspondent les vingt plus grandes valeurs qui entrent dans le calcul des estimations mais qui ne faisaient pas partie de la "vraie" liste des quarante plus grandes valeurs.

CHAMP	METHODE DU PLUS PROCHE VOISIN	METHODE DU "HOT DECK"
762	1	2
822	3	5
824	4	6
832	3	4
833	6	5
834	5	8
835	4	6
Moyenne	3.71	5.14

TABLEAU III (suite)

C. Rang (parmi les quarante plus grandes valeurs) de la plus grande nouvelle valeur introduite par l'imputation.

CHAMP	METHODE DU PLUS PROCHE VOISIN	METHODE DU "HOT DECK"
762	7	5
822*	1	1
824*	1	1
832	6	3
833	3	3
834	6	3
835	10	5
Moyenne	4.86	3.00

* La plus grande valeur qui entre dans le calcul de l'estimation de ce champ provient du même enregistrement, quelle que soit la méthode d'imputation, parce que le facteur d'accroissement est élevé et non parce que la valeur imputée est très grande.

ANNEXE A

Cette annnexe présente une liste des fichiers qui ont été préservés, de sorte que d'autres utilisateurs pourront avoir accès aux fichiers créés pendant l'exécution de cette étude.

SASKFES.DIRTY (vol = 024293) - 20% de non-réponse dans trois champs des enregistrements de la Saskatchewan. Contient tous les champs et les enregistrements de l'EEA (longueur : 2,400; 9,761 enregistrements)

SASKFES.ALL (vol = 046179) - 20% de non-réponse introduite dans trois champs. Contient seulement les enregistrements de la Saskatchewan et les champs nécessaires pour les imputations faites dans la présente étude. (longueur : 119; 3,644 enregistrements - sous-ensemble de SASKFES.DIRTY)

SASKFES.CLEANED.BYNN (vol = 062011) - Même structure que SASKFES.DIRTY. Fichier de données imputées par la méthode du plus proche voisin.

SASKFES.ALL.CLEAN (vol = 050922) - Même structure que SASKFES.ALL. fichier de données imputées par la méthode du plus proche voisin. (sous-ensemble de SASKFES.CLEANED.BYNN)

SASKFES.CLEANED.BYHD(vol = 011612) - Même structure que SASKFES.CLEANED.BYNN. Fichier de données imputées par la méthode du "hot deck".

ANNEXE B

Résumé des critères d'appariement des données manquantes et des valeurs imputées dans les deux méthodes d'imputation

Imputation	Code	CHAMPS touchés	Critères de sélection des valeurs imputées	
			Méthode du "hot deck"*	Méthode du plus proche voisin
135 Refus à la section des dépenses engagées pour le travail sur commande	Champ 760 = 9 (refus total de répondre à cette section)	762	Choisir 1 matrice sur 9 en fonction de 3 tranches de revenu (901) et de 3 intervalles de superficie totale (222)	Apparier aussi étroitement que possible en fonction du revenu (901) et de la superficie totale (222)
163 Refus à la section des dépenses engagées pour les intérêts	Champ 820 = 9 (refus total de répondre à cette section)	822 824	Choisir 1 matrice sur 6 en fonction de 6 tranches de revenu (901)	Apparier aussi étroiment que possible en fonction du revenu (901) et exactement selon la zone pédologique
169 Refus à la section des dépenses engagées pour le téléphone, l'électricité et le carbuarant	Champ 830 = 9 (refus total de répondre à cette section	832 833 834 835	Choisir 1 matrice sur 6 en fonction de 6 tranches de revenu (901)	Apparier aussi étroitement que possible en fonction du revenu (901)

* la méthode du "hot deck" divise également toutes les matrices en fonction de la zone pédologique; autrement dit, chaque imputation comporte implicitement un appariement avec la zone pédologique.